Flight Project

- Consider the following paper
    - https://www.dropbox.com/s/4rqnjueuqi5e0uo/TIST-Flight-Delay-final.pdf
- Half of the paper is dedicated to data preparation by preprocesing and opportunely *joining* complex datasets about flights and weather conditions
- CSV files are here https://www.dropbox.com/sh/iasq7frk6f58ptq/AAAzSmk6cusSNfqYNYsnLGlXa
- The report should detail each step, comment encountered difficulties and how these have been overcome.
- Data preparation and transformation are time-consuming, the join operation is rather complex
- Decision-tree, use Spark ML (do not implement it from scratch)
    - Use of other models are welcome
- First local machine (or Databricks) on a small amount of data, and then on the cluster on larger datasets.
- Scala
- ChatGPT ?
- Deadline : 2 weeks before the presentation
- Report : at least 20 pages, if many pages use 1 or more appendix for additional material